Geometric Noise for Locally Private Counting Queries

Lefki Kacem
University of Paris Saclay
France

Catuscia Palamidessi
INRIA and École Polytechnique
France

ABSTRACT

Local differential privacy (LDP) is a variant of differential privacy (DP) where the noise is added directly on the individual records, before being collected. The main advantage with respect to DP is that we do not need a trusted third party to collect and sanitise the sensitive data of the user. The main disadvantage is that the trade-off between privacy and utility is usually worse than in DP, and typically to retrieve reasonably good statistics from the locally sanitised data it is necessary to have access to a huge collection of them. In this paper, we focus on the problem of estimating the counting queries on numerical data, and we propose a variant of LDP based on the addition of geometric noise. Such noise function is known to have appealing properties in the case of counting queries. In particular, it is universally optimal for DP, i.e., it provides the best utility for a given level of DP, regardless of the side knowledge of the attacker. We explore the properties of geometric noise for counting queries in the LDP setting, and we conjecture an optimality property, similar to the one that holds in the DP setting.

CCS CONCEPTS

• **Security and privacy → Formal security models**; **Information flow control**; *Information-theoretic techniques*; *Privacy-preserving protocols*;

KEYWORDS

Local differential privacy, counting queries, geometric noise

ACM Reference Format:
Lefki Kacem and Catuscia Palamidessi. 2018. Geometric Noise for Locally Private Counting Queries. In *PLAS '18: The 13th Workshop on Programming Languages and Analysis for Security, Oct. 19, 2018, Toronto, ON, Canada.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3264820.3264827

1 INTRODUCTION

Our lives are growingly entangled with ubiquitous communication technologies and the limitless digital information they provide access to. The ways we relate to each other, work, travel, shop, or entertain ourselves are increasingly driven by mobile services. Most such services heavily rely on the collection and analysis of personal data, which are often generated and provided by the users themselves: tweeting about an event, browsing the World Wide Web, calling with a mobile phone, using a car navigation system, or paying with a credit card are examples of situations generating data. Service providers, web tracking platforms, mobile network operators, automotive manufacturers, or banking information systems can then gather substantial amounts of such data about millions of customers, at unprecedented accuracy level and at low cost.

This is the report version of a paper published in the proceedings of:
PLAS '18, October 19, 2018, Toronto, ON, Canada
https://doi.org/10.1145/3264820.3264827

While data-driven technologies provide undeniable benefits to individuals and society, the collection and manipulation of personal data has reached a point where it raises alarming privacy issues. Not only the experts, but also the population at large are becoming increasingly aware of the risks, due to the repeated cases of violations and leaks that keep hitting the headlines. Examples abound, from iPhones storing and uploading device location data to Apple without users' knowledge [1] to the popular Angry Birds mobile game being exploited by NSA and GCHQ to gather users' private information such as age, gender and location [5].

Until recently, the most popular and used data sanitization technique was anonymization (removal of names) or more sophisticated variants like k-anonymity [14] ensuring indistinguishability within groups of at least k people, and ℓ-diversity, ensuring a variety of values for the sensitive data within the same group [11]. Unfortunately, these techniques have been proved unable to provide an acceptable level of protection, as several works have shown that individuals in anonymized datasets can be re-identified with high accuracy, and their personal information exposed (see for istance [12, 13]).

In the meanwhile a new paradigm, *differential privacy* (DP) [8], has emerged and become extremely successful. In DP, the individual data are not directly accessible. Rather, the dataset is protected by an interface called *mechanism*. The mechanism then computes the answer on the basis of the information contained in the dataset and typically adds some controlled noise to the result, in such a way that the data of a single individual will have a negligible impact on the reported answer. DP has two important advantages with respect to other approaches: it is independent from the side-information of the adversary, thus a differentially-private can be designed without taking into account the context in which it will have to operate, and it is compositional, i.e., if we combine the information that we obtain by querying two differentially-private mechanisms, the resulting mechanism is also differentially-private.

In recent years, a variant of DP called local differential privacy (LDP) was proposed [7]. LPD is a distributed variant of differential privacy in which users obfuscate their personal data by themselves, in a differentially-private way, before they are collected. LPD implies DP on the resulting data collection, and has the same advantages of compositionality and independence from the side-information. Additionally, with respect to the centralized model, it has the advantages that it does not need to assume a trusted third party, and since all stored records are individually-sanitized, there is no risk of privacy breaches due to malicious attacks. Furthermore, each user can choose the level of privacy he wishes. LDP is having a considerable impact, especially after large companies such as Apple and Google have started to adopt it for collecting the data of their customers for statistical purposes [9].

Another variant of LDP, called d-privacy, has been proposed by [6]. The condition for d-privacy to be applicable is that the

domain of data is a metric space (the "d" in the name stands for "distance"), and the idea is to make advantage of the underlying topologic structure in order to improve the trade-off between utility and privacy. A well-known instance of d-privacy is the notion of *geo-indistinguishability* [4], which is obtained when the domain of data are locations, and the distance is the geographical distance. The main methods to obtain d-privacy are the extended laplacian and the extended geometric noise, applicable in the continuous and in the discrete cases, respectively. ("Extended" here means that the density function is defined on a generic metric space rather than on the reals.)

In this paper we consider the (truncated) geometric noise function, and the particular case of counting queries. The geometric noise has been proved *universally optimal* for counting queries [10], meaning that it is the mechanism providing the best utility for a given privacy level, and for all possible prior knowledge the adversary or the user may have. The notion of utility to which this result refers is the accuracy of the reported answer, measured (in the case of [10]) in terms of its expected distance from the true answer. We call it *punctual utility*.

When we deal with a collection of noisy data, however, there is another natural notion of utility, which is the capability to reconstruct the original distribution in order to make precise statistical analyses. The notion of d-privacy has been advocated in a recent work [3] as a variant of LDP that is able to provide a better trade-off between privacy and statistical utility than the standard LPD.

In this paper, we explore the properties of geometric noise for counting queries in the LDP setting, and we conjecture an universal optimality property, analogue to that of [10] but for LPD rather than DP, and for statistical utility rather than punctual utility.

2 PRELIMINARIES

In this section we recall some basic notions about probability distributions, DP, LPD, d-privacy, counting queries and geometric noise. We will consider only the case of discrete domains.

Given a generic set $\mathcal{X}$, a probability distribution p on $\mathcal{X}$ is a function $p : \mathcal{X} \to \mathbb{R}$ such that $\forall x \in \mathcal{X}\, p(x) \geq 0$ and $\sum_x p(x) = 1$. We will denote by $Distr(\mathcal{X})$ the set of all possible probability distributions on $\mathcal{X}$. For notational convenience, we will use p_x to denote $p(x)$.

2.1 Differential privacy

We will use D to denote a collection of data (dataset), $\mathcal{D}$ to denote the set of all possible datasets of the class of interest, and $\sim$ to represent the *adjacency relation* between datasets: $D \sim D'$ means that D and D' differ only for the value of a single record. We will assume that the dimension of the datasets in $\mathcal{D}$ is fixed, i.e. that each dataset in $\mathcal{D}$ contains exactly n records, for some n. Given a query $f : \mathcal{D} \to \mathcal{X}$, a (noisy) mechanism $\mathcal{K}$ for f is a probabilistic function which, for every D, gives a reported answer $y \in \mathcal{Y}$ with a certain probability distribution that depends on the true answer $x = f(D)$. The domain of the reported answers $\mathcal{Y}$ may coincide with $\mathcal{X}$ but not necessarily. We will use the notation $P[\mathcal{K}(D) = y]$ to denote the probability that $\mathcal{K}$ applied to D reports the answer y.

Then we say that $\mathcal{K}$ satisfies ε-DP, where ε is a non-negative real number denoting the level of privacy, if for every pairs of datasets D and D' such that $D \sim D'$, and for every $y \in \mathcal{Y}$, we have

$$P[\mathcal{K}(D) = y] \leq e^{\varepsilon}\, P[\mathcal{K}(D') = y]. \tag{1}$$

2.2 Local differential privacy

In LDP the idea is that the mechanism obfuscates directly the value of the data rather than the answer to a query. In this setting, we let $\mathcal{X}$ denote the set of all possible values (possibly tuples) for the data, and a mechanism $\mathcal{K}$ is a probabilistic function which, for every $x \in \mathcal{X}$, returns a reported value $y \in \mathcal{X}$ with a certain probability distribution that depends on the true value x. We will use the notation $P[\mathcal{K}(x) = y]$ to denote the probability that $\mathcal{K}$ applied to x reports the answer y.

A mechanism $\mathcal{K}$ provides ε-LPD if for every pair of input values $x, x' \in \mathcal{X}$, and for every measurable set S, we have

$$P[\mathcal{K}(x) = y] \leq e^{\varepsilon}\, P[\mathcal{K}(x') = y]. \tag{2}$$

2.3 d-privacy

In d-privacy, like in LDP, mechanism obfuscates directly the value of the data. The main difference is that the domain X is assumed to be a metric space, namely be endowed with a notion of distance $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^{\geq 0}$, where $\mathbb{R}^{\geq 0}$ is the set of non-negative real numbers.

A mechanism $\mathcal{K}$ provides ε-d-privacy if for every pair of input values $x, x' \in \mathcal{X}$, and for every $y \in \mathcal{Y}$, we have

$$P[\mathcal{K}(x) = y] \leq e^{\varepsilon\, d(x,x')}\, P[\mathcal{K}(x') = y]. \tag{3}$$

2.4 Counting queries

In DP, a counting query is a function $f : \mathcal{D} \to [0, n]$ such that $f(D)$ gives the number of records in D that satisfy a certain property. Here $[0, n]$ denotes the set of integers between 0 and n included.

In this paper, we will adopt a more abstract notion of counting query, suitable for LPD. Namely, we assume that $f : \mathcal{X} \to [0, n]$ associates a number $f(x) \in [0, n]$ to each element of $x \in X$.

For instance, $\mathcal{X}$ could be the set of records of a certain population, where each $x \in X$ contains information about a certain person, and f could be a function reporting, for example, the age (in years), or the number of children, or the monthly salary, etc. Namely, any function that encodes a query of the form "how many …", assuming that the result can be computed from x.

A mechanism $\mathcal{K}$ for f, in this context, associates to each value $i \in [0, n]$ a value $j \in [0, n]$ chosen randomly according to a certain probability distribution. We will denote by C_{ij} the probability that, on the element i, $\mathcal{K}$ reports j. Note that C_{ij} represent the conditional probability of i given j, hence the values C_{ij} form a stochastic matrix C (where C_{ij} is the element at the intersection of the i-th row and j-th column) such that $\forall i, j \in [0, 1]\, C_{ij} \geq 0$ and $\forall i \in [0, 1]\ \sum_j C_{ij} \geq 0$. From now on for notational simplicity we will use C rather than $\mathcal{K}$. We will also call C "mechanism" since it is a direct representation of $\mathcal{K}$.

2.5 Geometric mechanism

In the following, for simplicity we will use α to indicate $e^{-\varepsilon}$, where ε represents the level of privacy. Note that $0 < \alpha \leq 1$. The geometric mechanism (for a counting query) is represented by an infinite

matrix C with rows indexed by $[0, n]$ and columns indexed by $\mathbb{Z}$ (the set of integers), and whose elements are given by

$$C_{ij} = \frac{1-\alpha}{\alpha}^{|i-j|} \tag{4}$$

In order to avoid dealing with an infinite output domain, in general we consider the *truncated* version of the geometric mechanism. The idea is that the probability mass of every negative element is *remapped* in 0, and the probability mass of every element greater than n is *remapped* in n. The *truncated geometric mechanism* will be denoted by G and it is defined as

$$G_{ij} = \begin{cases} \frac{1}{1+\alpha}\alpha^{i} & j = 0 \\ \frac{1-\alpha}{1+\alpha}\alpha^{|i-j|} & 0 < j < n \\ \frac{1}{1+\alpha}\alpha^{|i-n|} & j = n \end{cases} \tag{5}$$

3 THE TRUNCATED GEOMETRIC AS A d-PRIVATE MECHANISM

In this section we investigate the properties of the truncated geometric mechanism. We start by observing that the truncated geometric is indeed an ε-d-private mechanism.

Proposition 3.1. *If $\mathcal{X}$ is the domain $[0, n]$ and d is the difference between integers, then G is a d-private mechanism on $\mathcal{X}$.*

Proof. The proof is immediate, taking into account that $\alpha = e^{-\varepsilon}$. Let $i, j, h \in [0, n]$, and assume first that $0 < j < n$. Then we have:

$$\begin{aligned} G_{ij} &= \frac{1-\alpha}{1+\alpha}\alpha^{|i-j|} \\ &\leq \frac{1-\alpha}{1+\alpha}\alpha^{|i-h|+|h-j|} \\ &= \alpha^{|i-h|}\frac{1-\alpha}{1+\alpha}\alpha^{|h-j|} \\ &= \alpha^{|i-h|}G_{hj}. \end{aligned}$$

where the second step is justified by the triangular inequality. For $j = 0$ and $j = n$ the proof is analogous. □

The following is an important property that will be used in the next section

Proposition 3.2. *The matrix G is invertible.*

Proof. (*Sketch*) Consider the relation between G and the geometric mechanism C defined by (4), and observe that the highest elements of C are all on the central diagonal (corresponding to the principal diagonal of G). Hence the rows of C are linearly independent. Since G is obtained from C by adding the columns of index $[-\infty, -1]$ to the column 0, and the columns of index $[n+1, \infty]$ to the column n, the rows of G are still linearly independent (G_{00} is still the highest element of column 0 and G_{nn} is still the highest element of column n). □

4 RECONSTRUCTING THE ORIGINAL DISTRIBUTION FROM A COLLECTION OF NOISY DATA

We consider now the following problem: Assume that we have a collection of N noisy data representing the result of the application of the geometric mechanism to the data of a certain population. Each datum (as well as each noisy datum) is a number in $[0, n]$, and let $\pi \in Distr([0, n])$ be the prior distribution on the original data. In other words, the set of original data is generated by a sequence of random variables $X_1, X_2, \ldots, X_N$ independent and identically distributed (i.i.d.), according to π. Let $p \in Distr([0, n])$ be the probability distribution determined by $X_1, X_2, \ldots, X_N$ (i.e., obtained by counting the frequencies of the result i in $X_1, X_2, \ldots, X_N$, for each $i \in [0, n]$. Namely, for how many h we have $X_h = i$).

To each of the results of $X_1, X_2, \ldots, X_N$ we apply the geometric mechanism G, thus obtaining a sequence of random variables $Y_1, Y_2, \ldots, Y_N$. Let $q \in Distr([0, n])$ be the probability distribution determined by $Y_1, Y_2, \ldots, Y_N$ (again obtained by counting the frequencies of j in $Y_1, Y_2, \ldots, Y_N$, for each $j \in [0, n]$).

The task we consider here is how best to reconstruct the original distribution π from q. To this purpose, we consider the following iterative procedure, which is inspired by the Bayes theorem. In the definition of this procedure, q_j represents the probability of j according to q, and analogously for $p_i^{(k)}$:

Definition 4.1. Let $\{p^{(k)}\}_k$ be the sequence of distributions defined inductively as follows:

$$\begin{aligned} p^{(0)} &= q \\ p_i^{(k+1)} &= \sum_j q_j \frac{p_i^{(k)}\alpha^{|i-j|}}{\sum_h p_h^{(k)}\alpha^{|h-j|}} \end{aligned}$$

The interest of the above definition relies in the following result:

Theorem 4.2. *Let $\{p^{(k)}\}_k$ be the sequence of distributions constructed according to Definition 4.1. Then:*

1. *The sequence converges, i.e., $\lim_{k\to\infty} p^{(k)}$ exists.*
2. *$\lim_{k\to\infty} p^{(k)}$ is the* Maximum Likelihood Estimator *(MLE) of p given q.*

Proof. (*Sketch*) The proof proceeds by showing that the algorithm to produce the sequence $\{p^{(k)}\}_k$ is an instance of the Expectation Maximization (EM) algorithm defined in [2], which proves the convergence to the Maximum Likelihood Estimator for all additive noise functions. □

We will denote by p^* the limit of the sequence $\{p^{(k)}\}_k$, i.e., $p^* \stackrel{\text{def}}{=} \lim_{k\to\infty} p^{(k)}$.

Theorem 4.2(2) means that for all possible distributions p', the probability that the distribution induced from the noisy data (sanitized with G) is q when the prior is p^* is higher than or equal to the same probability when the prior is p'.

Furthermore, as N increases, p and p^* approximate the prior π, as shown below. We first need the following lemma.

Lemma 4.3. *p^* is the unique fixed point of the transformation that generates the sequence $\{p^{(k)}\}_k$, namely, for every p',*

$$p'_i = \sum_j q_j \frac{p'_i\alpha^{|i-j|}}{\sum_h p'_h\alpha^{|h-j|}} \qquad \text{iff} \qquad p' = p^*$$

Proof. (*Sketch*) The if part is immediate by using the properties of the limit. The only if part follows from the fact that G is invertible. □

We are now ready to show the main result:

Theorem 4.4. *Let $\{p^{(k)}\}_k$ be the sequence of distributions constructed according to Definition 4.1. Then, as N grows, p and p^* approximate π. Namely:*

(1) $\lim_{N\to\infty} p = \pi$
(2) $\lim_{N\to\infty} p^* = \pi$

Note that the parameter N is implicit in the definition of p and p^*.

Proof. (*Sketch*)

(1) This part is standard and follows from the law of large numbers.
(2) This part follows from the fact that p^* is the MLE of p, from point (1) above, and from Lemma 4.3.

□

Finally, we give a characterization of p^* using G. For this we introduce the following notation: For a distribution $p \in Distr([0,n])$ and a matrix C indexed by $[0,n]\times[0,n]$, pC is the product of p (seen as a vector) and C. Namely, $(pC)_j = \sum_i p_i C_{ij}$. Furthermore, I will represent the identity matrix.

Proposition 4.5. *If $r = q\,G^{-1}$ is a probability distribution, then $p^* = r$.*

Proof. Thanks to Lemma 4.3, it is sufficient to prove that $r = qG^{-1}$ is a fixed point of the transformation. We have:

$$\begin{aligned}
\sum_j q_j \frac{r_i \alpha^{|i-j|}}{\sum_h r_h \alpha^{|h-j|}} &= \sum_j q_j \frac{r_i G_{ij}}{\sum_h (qG^{-1})_h G_{hj}} \\
&= \sum_j q_j \frac{r_i G_{ij}}{\sum_h (qI)_h} \\
&= \sum_j q_j \frac{r_i G_{ij}}{q_j} \\
&= \sum_j r_i G_{ij} \\
&= r_i
\end{aligned}$$

□

We conclude our investigation with a conjecture

Conjecture 4.6. *The truncated geometric mechanism with parameter $\alpha = e^{-\varepsilon}$ is the mechanism that gives the best approximation of the original distribution among the ones that are ε-d-private.*

The intuition is that the Arimoto algorithm for optimizing the trade-off between the mutual information and the distortion rate produces a mechanism of the form $\lambda_{ij} e^{-\varepsilon|i-j|}$, the mutual information is directly related to ε, and the distortion rate to the distance between the original distribution and the approximated one.

5 CONCLUSION

In this paper, we have investigated the properties of the truncated geometric mechanism in relation to the reconstruction from noisy data of the original distribution on the real data. We have provided an iterative algorithm to approximate the original distribution, and we have given a characterization of the fixed point in terms of the inverse of the matrix. Finally, we have conjecture the optimality of the truncated geometric mechanism with respect to the trade off privacy-statistical utility. In the future we intend to explore and try to prove this conjecture.

ACKNOWLEDGMENTS

The work of Catuscia Palamidessi has been partially supported by the ANR project REPAS.

REFERENCES

[1] April 20, 2011. 3G Apple iOS Devices Are Storing Users' Location Data. *The New York Times* (April 20, 2011).
[2] Dakshi Agrawal and Charu C. Aggarwal. 2001. On the Design and Quantification of Privacy Preserving Data Mining Algorithms. In *Proceedings of the Twentieth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS '01)*. ACM, New York, NY, USA, 247–255. https://doi.org/10.1145/375551.375602
[3] Mário S. Alvim, Konstantinos Chatzikokolakis, Catuscia Palamidessi, and Anna Pazii. 2018. Local Differential Privacy on Metric Spaces: Optimizing the Trade-Off with Utility. In *31st IEEE Computer Security Foundations Symposium, CSF 2018, Oxford, United Kingdom, July 9-12, 2018*. IEEE Computer Society, 262–267. https://doi.org/10.1109/CSF.2018.00026
[4] Miguel E. Andrés, Nicolás E. Bordenabe, Konstantinos Chatzikokolakis, and Catuscia Palamidessi. 2013. Geo-indistinguishability: differential privacy for location-based systems. In *Proceedings of the 20th ACM Conference on Computer and Communications Security (CCS 2013)*. ACM, 901–914. https://doi.org/10.1145/2508859.2516735
[5] James Ball. 2014. Angry birds and 'leaky' phone apps targeted by NSA and GCHQ for user data. *The Guardian* (January 27, 2014). www.theguardian.com/world/2014/jan/27/nsa-gchq-smartphone-app-angry-birds-personal-data.
[6] Konstantinos Chatzikokolakis, Miguel E. Andrés, Nicolás E. Bordenabe, and Catuscia Palamidessi. 2013. Broadening the scope of Differential Privacy using metrics. In *Proceedings of the 13th International Symposium on Privacy Enhancing Technologies (PETS 2013) (Lecture Notes in Computer Science)*, Emiliano De Cristofaro and Matthew Wright (Eds.), Vol. 7981. Springer, 82–102.
[7] John C. Duchi, Michael I. Jordan, and Martin J. Wainwright. 2013. Local Privacy and Statistical Minimax Rates. In *Proceedings of the 54th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*. IEEE Computer Society, 429–438. https://doi.org/10.1109/FOCS.2013.53
[8] Cynthia Dwork, Frank Mcsherry, Kobbi Nissim, and Adam Smith. 2006. Calibrating noise to sensitivity in private data analysis. In *In Proceedings of the Third Theory of Cryptography Conference (TCC) (Lecture Notes in Computer Science)*, Shai Halevi and Tal Rabin (Eds.), Vol. 3876. Springer, 265–284.
[9] Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. 2014. RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security (CCS)*, Gail-Joon Ahn, Moti Yung, and Ninghui Li (Eds.). ACM, 1054–1067. https://doi.org/10.1145/2660267.2660348
[10] Arpita Ghosh, Tim Roughgarden, and Mukund Sundararajan. 2009. Universally utility-maximizing privacy mechanisms. In *Proceedings of the 41st annual ACM Symposium on Theory of Computing (STOC)*. ACM, 351–360. https://doi.org/10.1145/1536414.1536464
[11] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. 2007. l-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 1, 1, Article 3 (2007). https://doi.org/10.1145/1217299.1217302
[12] A. Narayanan and V. Shmatikov. 2008. Robust De-anonymization of Large Sparse Datasets. In *2008 IEEE Symposium on Security and Privacy (S&P 2008)*. 111–125. https://doi.org/10.1109/SP.2008.33
[13] Arvind Narayanan and Vitaly Shmatikov. 2009. De-anonymizing Social Networks. In *Proceedings of the 30th IEEE Symposium on Security and Privacy (S&P 2009)*. IEEE Computer Society, 173–187. https://doi.org/10.1109/SP.2009.22
[14] Pierangela Samarati. 2001. Protecting Respondents' Identities in Microdata Release. *IEEE Trans. Knowl. Data Eng* 13, 6 (2001), 1010–1027. http://doi.ieeecomputersociety.org/10.1109/69.971193